



JRC SCIENCE FOR POLICY REPORT

On the Use of Web Data in Macroeconomic Forecasting

Monokroussos G.

2015

This publication is a Science for Policy report by the Joint Research Centre (JRC), the European Commission's science and knowledge service. It aims to provide evidence-based scientific support to the European policy-making process. The scientific output expressed does not imply a policy position of the European Commission. Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use which might be made of this publication.

JRC Science Hub

<https://ec.europa.eu/jrc>

JRC96610

EUR 28104 EN

PDF ISBN 978-92-79-61790-4 ISSN 1831-9424 doi:10.2791/829032

Luxembourg: Publications Office of the European Union, 2015

© European Union, 2015

Reproduction is authorised provided the source is acknowledged.

How to cite: George Monokroussos; On the Use of Web Data in Macroeconomic Forecasting; European Commission, Joint Research Centre; EUR 28104 EN; doi:10.2791/829032

All images © European Union 2015,

Title On the Use of Web Data in Macroeconomic Forecasting

Abstract

This report explores using web based information, together with big data in macroeconomic forecasting. Exploiting rich information sets has been shown to deliver significant gains in nowcasting and forecasting contexts, whereas indicators constructed using web data can lead to better nowcasts in the face of model and data uncertainty in real time, challenges which can be particularly relevant during business cycle turning points. An example illustrates the potential of this approach: for a period centered on the latest recession in the United States, that this approach has the potential to deliver particularly good real-time nowcasts of GDP growth.

On the Use of Web Data in Macroeconomic Forecasting

George Monokroussos

European Commission - Joint Research Centre

June 24, 2015

Introduction

Some of the most significant societal transformations that have taken place recently have been associated, either directly or indirectly, with the dramatic increase in the prevalence of the internet in our lives. One of the side-effects of this development is the abundance of data collected or constructed from internet activities of various sorts (in short, *web data*). While the implications of this expansion are still unfolding, and thus a full assessment and reckoning of pros and cons may be somewhat premature, there are certain characteristics of web data that are noteworthy and consequential (to varying degrees), especially when compared to data coming from more “traditional” sources. These characteristics span several dimensions, including volume and breadth, timeliness, frequency, quality and reliability in real time, heterogeneity and complexity.

A comparison between web data and traditional data could be quite favorable for the former in fields such as economics, in which it is difficult or even infeasible at times to conduct controlled experiments. One sub-field of economics for which using web data may hold particularly high promise is macroeconomics. Macroeconomics involves studying complicated and multifaceted phenomena at an aggregate level, for which no model may be adequate at all times, for instance around crisis periods or sharp contractions. This chal-

lenge is compounded by the well-known paucity of traditional macroeconomic information, to a large extent because of the relatively low frequency (typically quarterly or monthly) at which macroeconomic observations are available. Last, and certainly not least, we have the additional complications associated with real-time data and data revisions. For many of the key macroeconomic variables (e.g. GDP, unemployment, industrial production, etc.) observations can be highly inaccurate in real time, and thus susceptible to very substantial ex-post revisions. Of course, many of the most important policy decisions are taken in real time, and macroeconomic forecasts that are often needed for informed policy and business decisions are also made on the basis of information sets as they are available in real time, and without the benefit of hindsight. This issue has been recognized by influential contributions to the literature as being key for monetary policy and for macroeconomics in general; see, inter alia, Croushore and Stark (2001), Orphanides (2001), Giannone et al (2012).

Given the above challenges, web data may prove to be particularly valuable in economics in general, and in macroeconomics in particular, and not necessarily as a substitute, but rather as a complement to traditional data sources. To illustrate this point, the following two sections discuss a specific example, coming from Monokroussos (2014), where combining web data (from Google Trends in particular) with big data (coming from standard sources) shows how we can achieve substantial gains when forecasting US GDP. Much of the intuition underlying this is related precisely to the shortcomings discussed above of traditional macro data and modeling techniques, which can be partly overcome by the use of web data. The final section of this report outlines some issues with Google Trends and puts forth the case that information coming from the Europe Media Monitor of the JRC may be particularly useful in improving GDP and other macroeconomic forecasts.

Combining Web Data and Big Data in Macroeconomic Forecasting: An Example for the U.S. GDP

Obtaining accurate and timely forecasts of turning points in GDP growth is a central pre-occupation of macroeconometrics. This is quite justifiable, given the importance of the task for the private sector and for policy makers alike. However, good such forecasts can be difficult to obtain. For example, there is a substantial, decades-old literature (see, inter alia, Stekler (1972), Zarnowitz (1986), Loungani (2001), Lahiri and Wang (2013)) that presents ample evidence of such predictive failures spanning several countries, historical episodes of recessions, forecast horizons, and types of forecasters.

On the other hand, recent literature has demonstrated that it is possible to produce good *nowcasts* of GDP growth. For instance, the influential paper of Giannone, Reichlin and Small (2008) has convincingly demonstrated the gains to be made when nowcasting US GDP growth using dynamic factors to exploit information coming from a large data set.

In view of both of the above, one question that arises naturally is how well we can detect turning points in GDP growth in real time during the current quarter. Of course, this is a key question: Being able to accurately assess the present state of the economy, especially at the onset of recessions, as this is summarized by current-quarter GDP growth, is of central importance for the timely conduct of monetary policy, among other purposes. The most recent historical episode of the Great Recession is particularly telling.

The NBER (National Bureau of Economic Research) dates for the latest U.S. recession are December of 2007 for its peak and June of 2009 for its trough. One particularly consequential quarter in this recession was 2008 Q3, as it included events such as the collapse of Lehman Brothers, and it was also the first in a series of consecutive quarters with negative GDP growth. Of course, official estimates on 2008 Q3's negative growth only became available in the following (fourth) quarter. Similarly, the NBER called the recession on December 1, 2008. It is expected though that NBER recession announcements will not be the most timely

possible.

The same cannot be said however regarding the deliberations and decision making process of the monetary policy maker. Given likely lags in the monetary transmission mechanism, central banks have to rely heavily on macroeconomic forecasts. In particular, the Federal Reserve (the central bank of the United States) forecasts are generally perceived as being quite good (see, *inter alia*, Romer and Romer (2000) and Sims (2002)). However, the recently released minutes of the FOMC meetings covering the crucial period of the summer and fall of 2008 paint a picture of insufficient appreciation of (the extent of) the slowdown in real time and thus of the consequent policy risks and tradeoffs faced¹.

Given this, how would an approach that relies explicitly and exclusively on an econometric model fare? As was discussed above, the large-data, factor-based model of Giannone, Reichlin and Small (2008, henceforth GRS) is arguably at the peak of what we can achieve in GDP nowcasting contexts. Figure 1 shows the nowcasts of GDP growth obtained in real time using the model of GRS with historical vintages of close to 200 variables. As can be seen there, the model-based nowcasts do not turn negative until December of 2008 again. Furthermore, Giannone et al (2010) provide real-time nowcasts that turn negative late in the fall of 2008 and which are, however, more timely than either the respective Greenbook forecasts or the respective figures released by the Survey of Professional Forecasters.

In view of all of the above, one may tend to conclude that in the face of challenges such as model and data uncertainty, problems which can be particularly acute in real time during turning points, there may be little more that we can do. We may have to settle for nowcasts which are inferior around such turning points (when they are arguably needed the most) than at other times.

A central contribution of Monokroussos (2014) is to show that such pessimistic assess-

¹Matthew O'Brien scrutinized the recently released minutes of the June 24-25, August 5, and September 16, 2008 FOMC meetings in terms of keyword counts (e.g. frequency of the word "inflation" vs. "unemployment" or "systemic risks/crises") as well as in terms of specific statements by participants in the FOMC meetings and provides a substantial series of evidence along these lines (O'Brien (2014)). The Greenbook forecasts for 2008 Q3 GDP growth associated with these three FOMC meetings were all positive and indeed close to 1%.

ments do not have to be true any longer. Given the contemporary prevalence of the internet and internet search engines, we now have forecasting tools at our disposal that were not available until recently. GDP growth turning from positive to negative typically entails a widespread slowdown in economic activity. Workers, investors, employers, etc. who experience a change in their conditions that is associated with the slowdown are more likely to conduct internet searches using keywords related to the slowdown than at other times. Internet-based services such as *Google Trends* construct normalized time series indices reflecting the relative volume of such keyword searches. Such measures may provide a valuable gauge of the economy in real time as they have the potential to capture widespread changes in conditions in a timely manner and are also not subject to revisions and real-time inaccuracies like many of the more traditional variables are.

Monokroussos (2014) proposes and estimates a new bayesian model for a policy maker or any forecaster in general who, rather than operate within the confines of a traditional data set, instead “listens to hoi polloi” too, that is, lets her prior beliefs be influenced by internet search popularity measures. Forecasts emerge from posterior estimates that reflect both such prior beliefs and information coming from large traditional data sets. Dynamic factors are used to capture the collinearities and summarize these large data sets in a parsimonious way without throwing away information.

The main empirical result is that for a time period centered around the Great Recession, this bayesian factor-based nowcasting approach with popularity priors delivers a more timely detection of the 2008 Q3 turning point than all of the other alternatives discussed above (as is illustrated in Figure 3). Furthermore, it achieves a substantially better outcome regarding this consequential turning point while its nowcasts for the rest of the time are (at least) as good as the ones obtained from alternative approaches in real time.

More specifically, the model estimated consists of the following three equations:

$$x_{t|\tau_t} = \mu + \Lambda F_t + \xi_{t|\tau_t} \tag{1}$$

where we assume that time is measured in months t , and days τ_t , that is business days of month t when new data releases on one or more variables in our data set become available, x_i with $i = 1, \dots, N$ are the “traditional” variables of the big data set (close to 200 variables in our case), that contain information that is potentially useful for the macroeconomic aggregate we wish to forecast, μ is an $N \times 1$ vector of constants, and $\xi_{t|\tau_t}$ is an $N \times 1$ vector of idiosyncratic error terms which are Gaussian white noises, and are also cross-sectionally orthogonal. That is, $E(\xi_{t|\tau_t} \xi'_{t-s|\tau_t}) = 0$, for all $s > 0$, and all τ, t , and $E(\xi_{t|\tau_t} \xi'_{t|\tau_t}) = \Sigma^\xi = \text{diag}(\sigma_{\xi_1}^2, \dots, \sigma_{\xi_N}^2)$. Λ is an $N \times r$ vector of factor loadings, and the $r \times 1$ vector $F_t = (F_{1t}, \dots, F_{rt})'$, is the set of factors that capture the key information of the big data set, and are orthogonal to the error terms.

Regarding these factors, the literature commonly employs dynamic specifications that allow for inertia and that can capture intertemporal relationships among variables during the business cycle, and here we follow this standard paradigm by specifying a first order vector autoregression for the factors:

$$F_t = AF_{t-1} + \zeta_t \quad (2)$$

where A is an $r \times r$ coefficient matrix with all the roots of $\det(I_r - Az)$ lying outside the unit circle, ζ_t is an $r \times 1$ vector of “common shocks”, Gaussian white noises that are independent from the idiosyncratic error terms and with $E(\zeta_t \zeta'_t) = \Sigma^\zeta$ and where F_t, F_{t-1} are the current and last month’s factors.

Finally the model is complete with a so-called “bridge equation” that delivers forecasts or nowcasts of the macro aggregate in question, y , as a function of the factors:

$$y_{t+h|\tau_t} = \alpha + \beta' F_{t+h} + \varepsilon_{t+h|\tau_t} \quad (3)$$

where y is the macroeconomic variable we seek to forecast (GDP in this case) α is scalar and β is an $r \times 1$ vector of coefficients, h is the forecast horizon, thus $h = 0$ (for nowcasts) or

higher, and $\varepsilon_{t+h|\tau_t}$ is also a Gaussian white noise with variance σ_ε^2 .

The above model is estimated by Bayesian MCMC methods. The reader is referred to Monokroussos (2014) for further details regarding the model and its estimation including the MCMC algorithm, as well as a discussion of the challenges encountered when estimating such models, including degrees of freedom, missing data, and mixed frequencies in the data.

Nowcasting US GDP During the “Great Recession”

We employ the model and estimation algorithm described in the previous sections to nowcast US GDP growth, with a focus on the latest recession. As is well known, the U.S. Bureau of Economic Analysis does not release its estimates of current quarter GDP growth until next quarter; it releases a preliminary (“Advance”) estimate towards the end of the first month of the following quarter, and then it updates this figure one and two months later (“Second” and “Third” estimates, respectively). This constitutes a significant lag, especially for monetary policy purposes. Thus the task of nowcasting GDP growth (as well as other aggregates subject to similar lags in releases), of obtaining that is current-quarter GDP growth estimates during the current quarter, becomes particularly relevant, and the attention that the burgeoning nowcasting literature has been receiving is well justified.

We adopt here the perspective of a monetary policy maker, or of a professional forecaster, who needs to assess the current state of the economy in real time as accurately as possible and given all the available information. Assuming a formal modeling approach is adopted, it is arguably desirable to consider frameworks that, as discussed earlier, can handle large data sets, with jagged edges, and possibly mixed frequencies. Furthermore, if we hope to provide a realistic depiction of the nowcasting environment, with the regular influx of possibly inaccurate real-time information, we should be employing real-time data, rather than revised series that only became available ex-post. So, in this section we use a real-time data set - indeed a series of weekly, Friday jagged-edge data sets reflecting all the updates (new

observations or revisions of existing observations) that took place during the week. There are close to 200 (mostly monthly) macroeconomic variables, including monetary aggregates, prices, employment statistics, survey data, housing, banking balance sheet figures, etc. They include observations starting in January of 1982, with the weekly real time vintages starting in March of 2005

At the core of our approach is the realization that when aggregate growth turns from positive to negative, this typically reflects a widespread slowdown in economic activity, which thus affects a wide spectrum of people, either directly or indirectly. Given the prevalence of the internet in our time, many people may turn to internet search engines in an attempt to better understand the economy and their changing conditions. They may search related keywords in higher volumes than at other times. Tools such as Google Trends measure the volume of keyword-based searches and have begun compiling and making publicly available normalized time series indices based on the volume of searches. There are already interesting forecasting/nowcasting applications using such indices including Askitas and Zimmermann (2009), D’Amuri and Marcucci (2010) and Scott and Varian (2014a,b). The growing interest is certainly justified given that such indices have several key features, including that they reflect relative, and not absolute, volumes of searches, that they are based on real-time information, and that they, for certain keywords, may be able to capture widespread changes in financial and economic conditions in a timely manner, without being plagued by real-time inaccuracies and revision issues.

Helpful keywords in the present context would lead to Google indices that stay relatively flat at times when models such as the ones discussed above nowcast well, and that spike up when we do not nowcast well, primarily in the fall of 2008. An example of such a helpful keyword is “recession”, whose Google index is shown in Figure 2. We can notice there that it spikes up not only in the fall of 2008, but also in the period around the end of 2007 and the beginning of 2008 (which includes the NBER peak).

A policy maker/nowcaster, well aware of the above, may conclude that “listening to the

people too” in addition to consulting information sets based on “traditional” large data sets, is a promising nowcasting approach, with the potential to deliver superior results around turning points. Our approach allows such a nowcaster to take both into account. She lets internet search popularity measures such as Google Trends inform her prior beliefs. Thus her nowcasts reflect both such “popularity priors” and information coming from the dynamic factors and the large data sets as discussed above.

This approach can be implemented in various ways. In the example that follows, we propose one such possibility, using the “recession” keyword, and a simple, conservative approach towards constructing priors. The model and all the priors remain the same as above (see Table 1), with the only one changing being that of α , the intercept of equation 3. Specifically, the mean of the prior of α is set to the post-WWII recession (duration-weighted) average. Its prior variance follows the schedule below:

$$Variance(\alpha) : \left\{ \begin{array}{ll} 1 & \text{when } GRI \text{ is between } 0 \text{ and } 9 \\ 0.9 & \text{when } GRI \text{ is between } 10 \text{ and } 19 \\ \vdots & \vdots \\ 0.1 & \text{when } GRI \text{ is between } 90 \text{ and } 99 \\ 0.05 & \text{when } GRI \text{ is } 100 \end{array} \right. \quad (4)$$

Quite clearly, this postulates an increasing level of certainty that we are indeed in a recession as *GRI*, the *Google Recession Index* (Figure 7) increases.

The resulting real-time nowcasts obtained with this model are provided in Figure 3, and are contrasted with those coming from other models discussed above. The key finding is that the Bayesian model with the above popularity prior delivers a more timely recognition of the turning point in GDP growth in the fall of 2008, and substantially so, as the nowcast now turns negative by the end of Q3/beginning of Q4 of 2008². It is interesting to note here if

²This is the nowcast we could have obtained in the morning of October 1, 2008, and thus reflects information available up to Quarter 3. Recall that in this application we produce one nowcast per month, in the very beginning of each month. Of course, our methodology can generate updated nowcasts many times each month, any time new information becomes available.

the benchmark by which nowcasts are to be judged is advance GDP releases (depicted in Figure 3), then the nowcast in question (produced on October 1, 2008) undershoots by a significant amount. However, these estimates can be subject to significant revisions themselves, especially around turning points, while this is not a concern with the GRI. Figure 4 depicts revised GDP estimates; the revised estimate corresponding to 2008 Q3 is indeed much closer to our nowcast using popularity priors.

Furthermore, the improvement discussed in the previous paragraph, which can be quite consequential for monetary policy purposes in real time, does not come at the cost of deteriorated performance at other times, when compared to either the GRS nowcasts or the Bayesian nowcasts with the uninformed priors. Indeed, a policy maker with popularity priors would have actually been closer to the “truth” more often during the three years in question (56.3% of the time closer to advance GDP estimates and 63% of the time closer to revised GDP estimates) than she would have been had she relied exclusively on the traditional data. Monokroussos (2014) provides further information on this comparison and results on statistical significance issues.

Taking all of the above into account, we can conclude that the Bayesian nowcasting model with popularity priors achieves a substantially better outcome regarding the consequential turning point of the fall of 2008, while its nowcasts for the rest of the time are (at least) as good as the ones obtained from alternative approaches in real time.

The above is indicative of what can be achieved using this approach, which is quite flexible, as it can accommodate other specifications as well, as discussed in Monokroussos (2014).

Some Issues with Google Trends and the JRC's Europe Media Monitor

As the above example illustrates, web data have significant potential in economic forecasting contexts. This, as well as the fact that the related literature in economics is nascent, indicate a very promising avenue for academic research and policy work. The potential of resources such as Google Trends needs to be investigated more. Furthermore, other resources for useful web-based information need to be explored. Close to home, one particularly important such example is the JRC's Europe Media Monitor (EMM).

EMM may be a resource that compares favorably to Google Trends, particularly for forecasting tasks such as the one described above, for various reasons: Google trends data are available in the public domain only to a limited degree: indeed, for longer time intervals, we get only one observation per month. Google-trends observations that are publicly available are not based on the entire volume of Google searches on the relevant keyword(s), but rather on a random sample drawn from that volume of searches. Additionally, there are some further restrictions and limitations to the output obtained from Google trends when results are sought on the basis of not only keyword(s), but also geographical location(s).

The Europe Media Monitor compares favorably on all these fronts. It monitors media in real time, every ten minutes, and as such it can provide keyword-based indices at a much higher frequency than monthly or weekly. It monitors around 4000 news sources worldwide, processes around 200000 (new) news articles every day, and scans the web in 60 languages, thus having the capability of delivering high-quality country-specific information. Furthermore, it calculates article "tonality" indices, thus providing rich information not only in terms of frequency, but also in terms of heterogeneity.

These features allow for bayesian modeling strategies that combine traditional and web-based information such as the one described above, but with the potential of employing more sophisticated priors than the arguably crude prior of the example discussed above. They are

thus of high promise in economic forecasting contexts, and modeling approaches that take full advantage of EMM's capabilities ought to be explored.

References

- Askatas, N. , and K. Zimmermann (2009): "Google Econometrics and Unemployment Forecasting," *IZA Discussion Paper* No. 4201.
- Croushore, D. and T. Stark (2001): "A Real-Time Data Set for Macroeconomists," *Journal of Econometrics* 105(1), 111-130.
- D'Amuri, F. and J. Marcucci (2010): "Google it! Forecasting the US Unemployment Rate with a Google Job Search Index," *Nota di Lavoro della Fondazione Eni Enrico Mattei* 31.2010.
- Giannone, D., J. Henry, M. Lalik, M. Modugno (2012): "An Area-Wide Real-Time Database for the Euro Area," *Review of Economics and Statistics*, 94(4), 1000-1013.
- Giannone, D., M. Modugno, L. Reichlin, and D. Small (2010): "Nowcasting in Real-Time," *mimeo*, presented at the *16th International Conference on Computing in Economics and Finance, Society for Computational Economics*, London, UK. July 13.
- Giannone, D., L. Reichlin, and D. Small (2008): "Nowcasting: The Real-Time Informational Content of Macroeconomic Data," *Journal of Monetary Economics* (55), 665-676.
- Lahiri, K., and J.G. Wang (2013): "Evaluating Probability Forecasts for GDP Declines Using Alternative Methodologies," *International Journal of Forecasting* 29(1), 175-190.
- Loungani, P. (2001): "How Accurate are Private Sector Forecasts? Cross-country Evidence from Consensus Forecasts of Output Growth," *International Journal of Forecasting* 17(3), 419-432.

- Monokroussos, G. (2014): “Nowcasting in Real Time Using Popularity Priors,” *mimeo*, *European Commission, Joint Research Centre*.
- O’Brien, M. (2014): “How the Fed Let the World Blow up in 2008,” *The Atlantic*, February 26, 2014.
- Orphanides, A. (2001): “Monetary Policy Rules Based on Real-Time Data,” *American Economic Review* 91(4), 964-985.
- Scott, S.L. and H.R. Varian (2014a): “Predicting the Present with Bayesian Structural Time Series,” *Int. J. Mathematical Modeling and Numerical Optimisation*, Vol. 5, Nos. 1/2.
- Scott, S.L. and H.R. Varian (2014b): “Bayesian Variable Selection for Nowcasting Economic Time Series,” *mimeo*.
- Stekler, H. (1972): “An Analysis of Turning Point Forecasts,” *American Economic Review* 62 (4), pp. 724-729.
- Zarnowitz, V. (1986): “The Record and Improvability of Economic Forecasting,” *NBER Working Paper 2099*.

Figure 1: Real-Time Nowcasts and GDP Growth

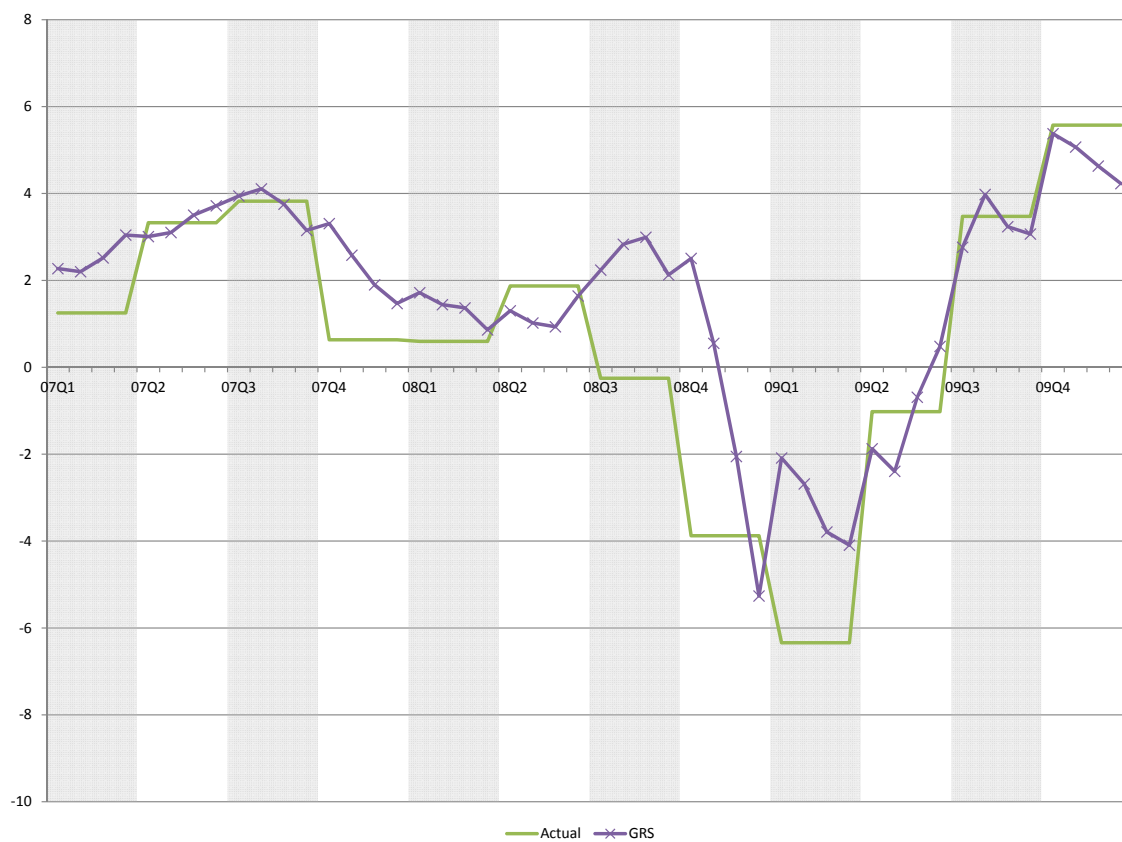


Figure 2: Google Searches for Keyword “Recession”

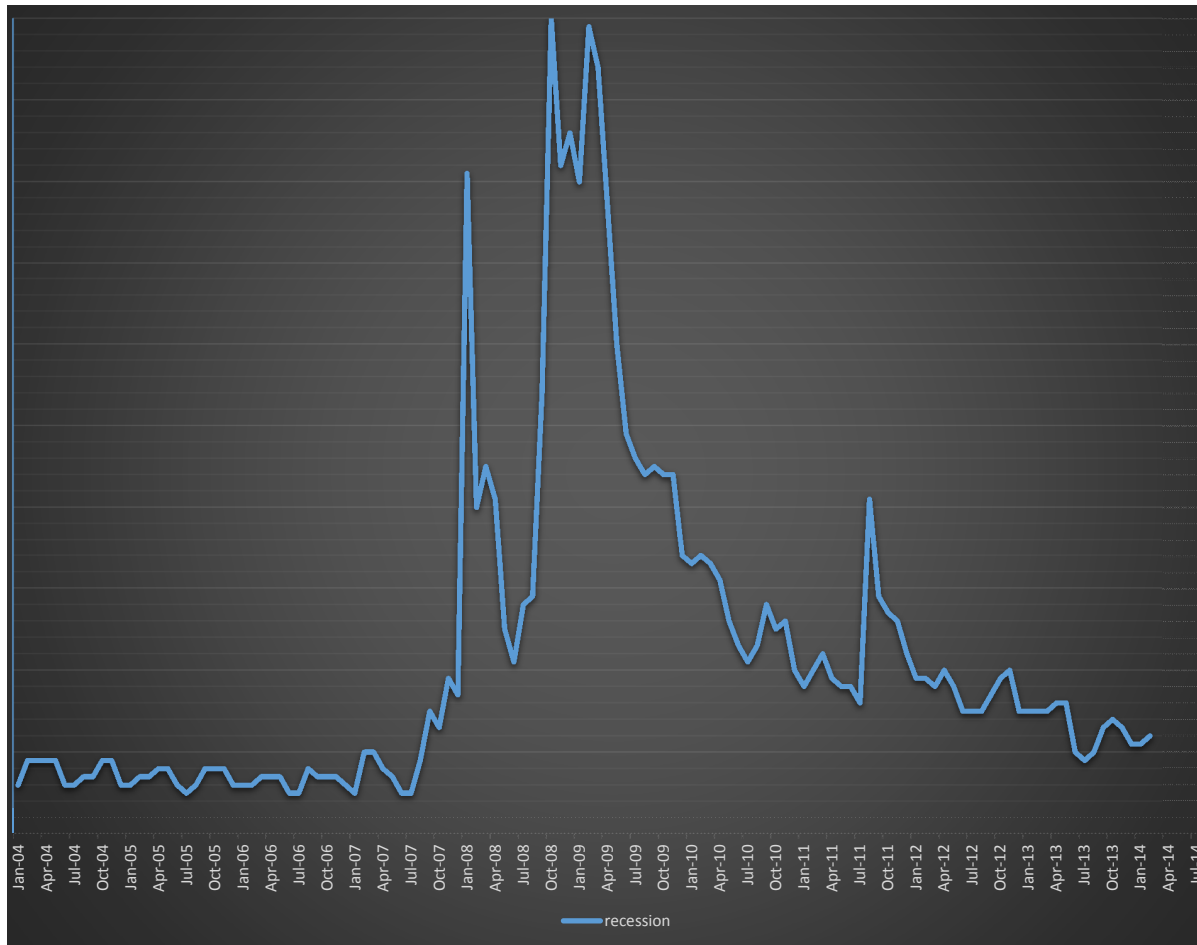


Figure 3: Real-Time Nowcasts with Popularity Priors

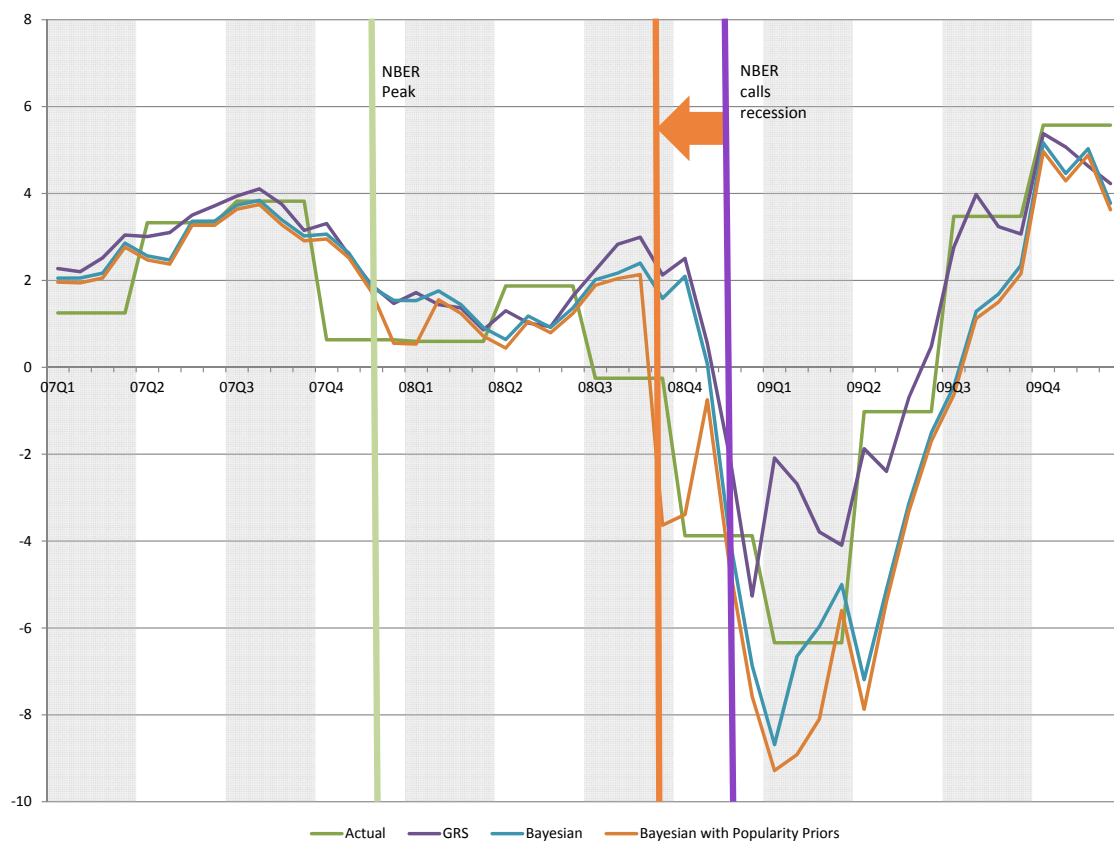
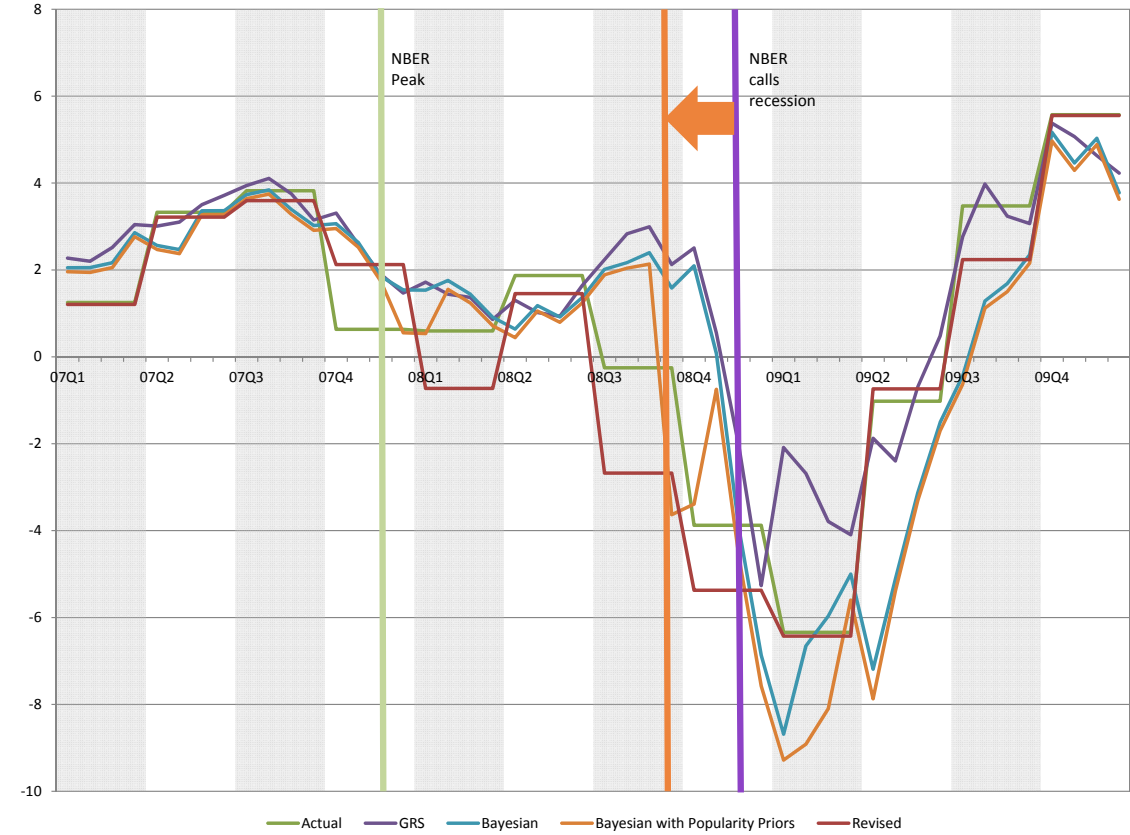


Figure 4: Real-Time Nowcasts with Popularity Priors and Revised Estimates of GDP Growth



Europe Direct is a service to help you find answers to your questions about the European Union
Free phone number (*): 00 800 6 7 8 9 10 11
(*) Certain mobile telephone operators do not allow access to 00 800 numbers or these calls may be billed.

A great deal of additional information on the European Union is available on the Internet.
It can be accessed through the Europa server <http://europa.eu>

How to obtain EU publications

Our publications are available from EU Bookshop (<http://bookshop.europa.eu>),
where you can place an order with the sales agent of your choice.

The Publications Office has a worldwide network of sales agents.
You can obtain their contact details by sending a fax to (352) 29 29-42758.

JRC Mission

As the science and knowledge service of the European Commission, the Joint Research Centre's mission is to support EU policies with independent evidence throughout the whole policy cycle.



EU Science Hub

ec.europa.eu/jrc



@EU_ScienceHub



EU Science Hub - Joint Research Centre



Joint Research Centre



EU Science Hub

